Rapport Projet Fairness

Baptiste PRAS

6 avril 2025

Table des matières

1 Introduction

1.1 Dataset et Objectifs

Le dataset utilisé $Chest\ X-ray\ NIH$ 14 est composé d'images de radiographies de thorax réalisés à l'hôpital. Nous possédons un fichier *metadata.csv* où sont retranscrites pour chaque image des informations sur les patients telles que leur âge, leur genre, la vue de l'image, l'ID du patient, l'ID de l'image et la maladie de laquelle souffre le patient. Il y a 14 maladies possibles, et il est également possible que le patient soit sain.

Le but ici est d'étudier et de corriger les biais possibles lors de la prédicition sur une image. Le but de la prédiction est simplement de prédire si le patient est malade ou sain. On peut alors estimer que cela sert juste de premier avis médical, et qu'un patient jugé malade subira des examens complémentaires, alors qu'un patient jugé sain sera renvoyé à la maison. Il est donc intéressant de maximiser les $Vrais\ Positifs$, quite à avoir plus de $FauxPositifs$ (dans la limite du raisonnable) pour éviter de ne pas soigner un patient qui en aurait besoin. On définira donc à partir de maintenant la prédiction *malade* comme étant la prédiction *positive*.

1.2 Méthodes et Mise en Place du Projet

Pour mettre en place le projet, nous avons décidé d'utiliser un fichier *main.py* où créer et utiliser toutes les fonctions nécessaires à la réalisation du projet. Le fichier $train\_classifier.ipynb$ sert seulement à créer un modèle avec l'appel à la fonction $train\_classifier()$ et à évaluer ce modèle avec l'appel à la fonction $pred\_classifier()$.

Les fichiers $versions\_[nom\ du\ biais].txt$ permettent de stocker les différentes métriques et performances de chaque modèle utilisé. Le dossier *plots* contient lui tous les plots générés pour évaluer les modèles graphiquement.

2 Préparation et Analyse des Données

2.1 Préparation

Pour préparer les données, nous avons déjà réduit le nombre d'images avec la fonction fournie dans *data_manipulation.ipynb* pour pouvoir entraîner nos modèles et faire les prédictions en un temps raisonnable (avant ce pré-traitement, cela prenait environ 50 minutes de faire un entraînement, et 5 à 10 minutes de faire les prédictions).

Par la suite, nous avons d'abord vérifié dans *metadata.csv* que toutes les données étaient correctes. Via la fonction *check_data*(), nous avons vérifié qu'il n'y avait pas de valeur manquante et que toutes les valeurs étaient cohérentes. En particulier pour l'âge, nous avons vérifié qu'il n'y avait pas de personne avec un âge particulièrement improbable. Après analyse, nous avons pu écarter cette hypothèse en trouvant que tous les âges étaient compris entre 2 et 91.

2.2 Analyse

Pour trouver les biais, nous nous sommes concentrés sur 2 catégories pouvant être discriminées : *l'âge* et *le genre*. Les autres caractérisques comme la vue de l'image ou le nombre de follow-ups ne sont pas des caractérisques qui prêtent à discriminer une catégorie de personnes en particulier et n'ont donc pas besoin d'être étudiées.

Pour étudier les biais, nous avons créé dans *metadata.csv* une colonne $Labels$, créé via la fonction *initialize_labels*() afin d'avoir l'information sur un patient s'il est sain ou malade. La valeur vaut *sain* si dans la colonne $Finding\ Labels$ se trouve l'information $No\ Finding$, et vaut *malade* sinon.

Les métriques utilisées sont le Taux de Vrais et Faux positifs en fonction de l'attribut sensible et la différence absolue de TPR et FPR entre les groupes, le but étant de minimiser ces différences, et en priorité pour le TPR comme précisé en introduction. Le calcul de ces métriques se fait via la fonction *equalized_odds*().
Le premier classifier, utilisé sans aucune correction des biais, avec donc tous les poids initialisés à 1.0, donnait les performances suivantes :

$$balanced\_accuracy = 0.671 \mid accuracy = 0.687$$

2.2.1 Biais sur l'âge

Pour vérifier les biais sur l'âge, nous avons d'abord dû créer une classification binaire sur l'âge. Pour cela, nous avons créé la colonne $Age\ Category$ avec la fonction *initialize_age_category*(), qui classifie les personnes en $Jeune$ ($\leq 50\ ans$) ou $Vieux$ ($> 50\ ans$).

Nous avons ensuite pu réaliser une analyse bivariée sur les colonnes $Age\ Category$ et $Labels$ avec la fonction *bivariate_analysis*(). Cette fonction affiche d'abord la matrice de répartition des différentes valeurs de la colonne $Age\ Category$ en fonction des différentes valeurs de la colonne $Labels$. En résumé, nous avons ici affiché le nombre de jeunes malades, de jeunes sains, de vieux malades et de vieux sains.

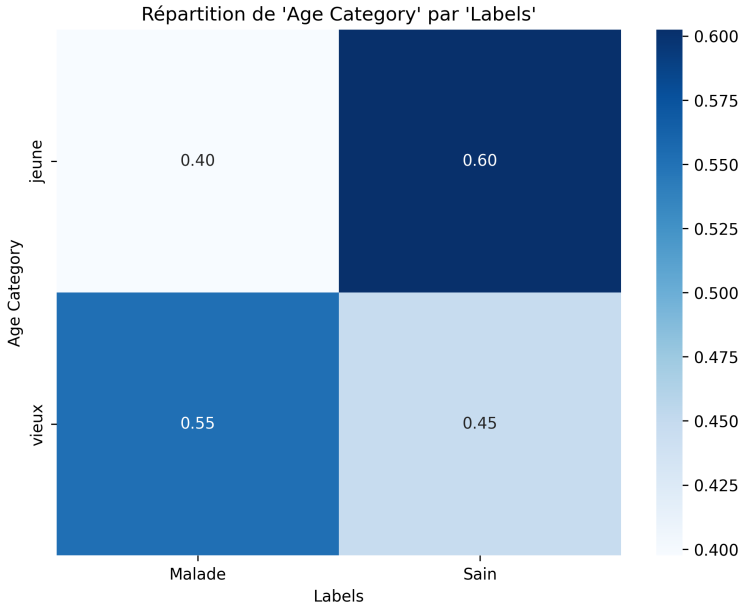

FIGURE 1 – Répartition des individus par âge et label

Enfin, nous avons analysé si la catégorie d'âge était concernée par un fort biais ou non avec la fonction de scipy $chi2\_contingency()$ qui renvoie les valeurs χ^2 et $p-value$. χ^2 mesure la différence entre les prédictions attendues et les prédictions obtenues. Plus la différence est grande, plus χ^2 est grand et donc plus les données prédites diffèrent de la distribution attendue. Cela peut alors indiquer un biais important. $p-value$ est la probabilité d'accepter l'hypothèse "La différence observée est dûe au hasard", donc plus p tend vers 0, plus il est improbable que ce qu'on observe soit dû au hasard, suggérant alors un potentiel biais. En outre, plus χ^2 est grand, plus $p-value$ tendra vers 0.

$$\chi^2 = 35.22 \mid p-value = 0.0000$$

Avec de telles valeurs, il semble fortement probable qu'il existe un biais sur l'âge qui devra être étudié plus en profondeur et corrigé par la suite.

Nous avons ensuite regardé la répartition des Vrais Positifs et Faux Positifs par âge grâce aux fonctions $show\_tpr()$ et $show\_fpr()$ pour voir plus en détail le potentiel biais.

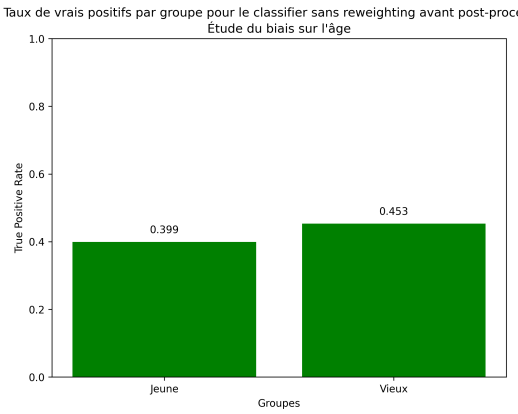

FIGURE 2 – Proportion de Vrais Positifs en fonction de l'âge

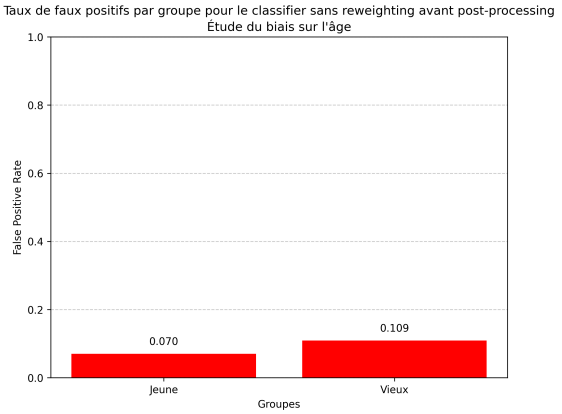

FIGURE 3 – Proportion de Faux Positifs en fonction de l'âge

$$Différence\ TPR : 0.054 \mid Différence\ FPR : 0.039$$

On peut remarquer que le modèle sans correction des biais a une tendance à prédire plus souvent un cas positif pour une personne âgée, en étant meilleur sur les Vrais Positifs pour ces dernières par rapport aux jeunes, mais moins bon sur les Faux Positifs.

2.2.2 Biais sur le genre

Pour vérifier les biais sur le genre, nous avons réalisé une analyse bivariée sur les colonnes *Patient Gender* et *Labels* avec la fonction *bivariate_analysis*(). Nous avons ici donc affiché le nombre de femmes malades, de femmes saines, d'hommes malades et d'hommes sains.

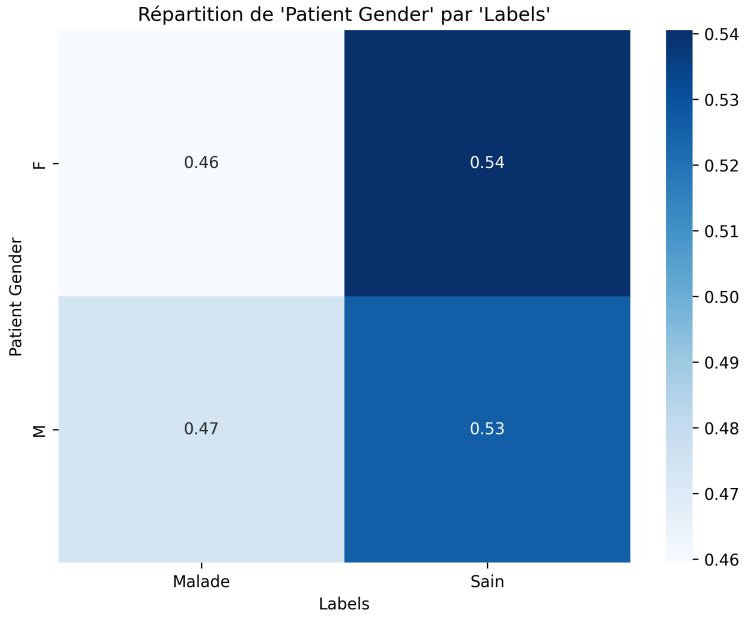

FIGURE 4 – Répartition des individus par genre et label

Enfin, nous avons analysé si le genre était concernée par un fort biais ou non avec la fonction de scipy *chi2_contingency*() qui renvoie les valeurs χ^2 et $p - value$.

$$\chi^2 = 0.27 \mid p - value = 0.6015$$

Avec de telles valeurs, il semble qu'il n'y ait pas ou peu de biais sur le genre, cela devra tout de même être étudié plus en profondeur et potentiellement corrigé par la suite.

Nous avons ensuite regardé la répartition des Vrais Positifs et Faux Positifs par genre grâce aux fonctions *show_tpr*() et *show_fpr*() pour voir plus en détail le potentiel biais.

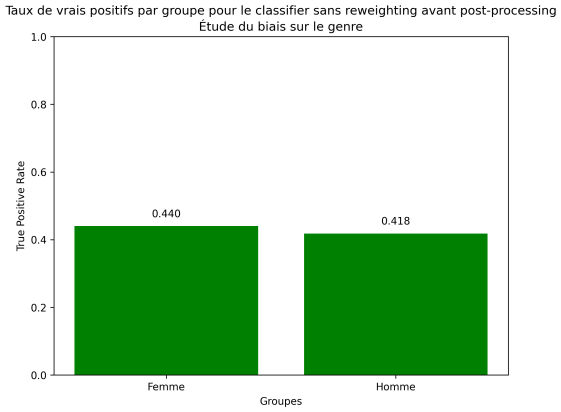

FIGURE 5 – Proportion de Vrais Positifs en fonction du genre

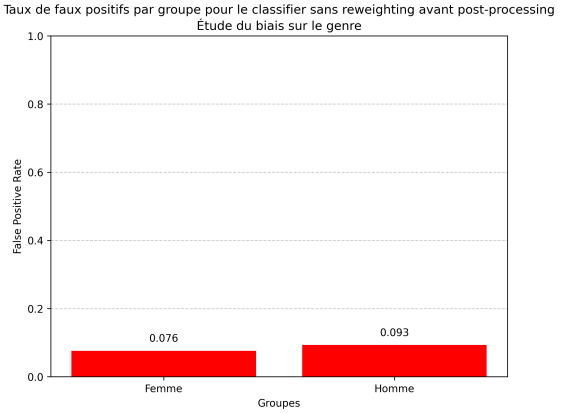

FIGURE 6 – Proportion de Faux Positifs en fonction du genre

$$Différence\ TPR : 0.022 \mid Différence\ FPR : 0.017$$

On peut remarquer que le modèle sans correction des biais produit de meilleurs résultats pour les femmes, que ce soit en Vrais Positifs comme en Faux Positifs.

2.2.3 Biais sur l'âge et le genre

Enfin, l'âge et le genre étant potentiellement biaisés, nous avons voulu faire une analyse sur les deux combinés. Pour cela, nous avons d'abord dû créer la colonne *Age+Gender* avec la fonction *initialize_combined_age_gender*(), qui classifie les personnes selon leur âge et genre. On obtient donc les labels *F_jeune* (jeune femme), *F_vieux* (vieille femme), *H_jeune* (jeune homme) et *H_vieux* (vieil homme).

Nous avons ensuite réalisé une analyse bivariée sur les colonnes $Age + Gender$ et *Labels* avec la fonction *bivariate_analysis*(). Nous avons ici donc affiché le nombre de jeunes femmes malades, de vieilles femmes malades, de jeunes femmes saines, de vielles femmes saines, de jeunes hommes malades, de vieils hommes malades, de jeunes hommes sains et de vieils hommes sains.

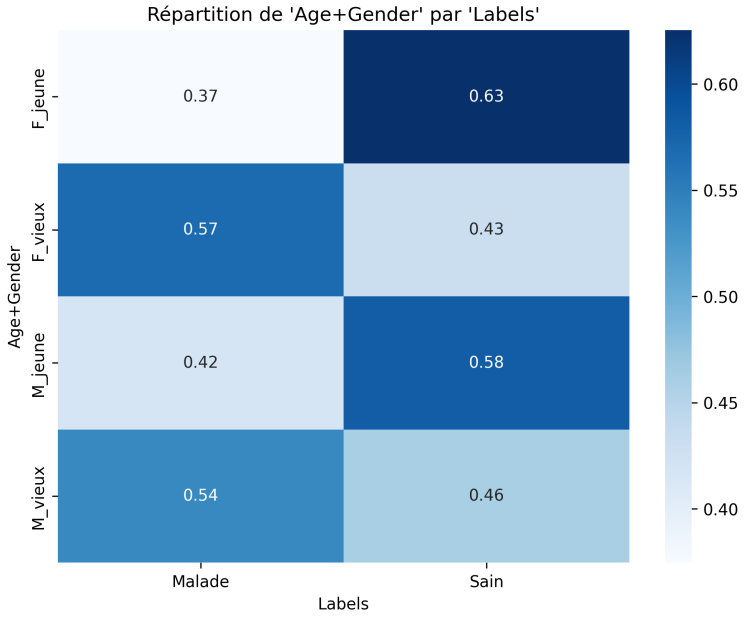

FIGURE 7 – Répartition des individus par âge+genre et label

Enfin, nous avons analysé si le genre était concernée par un fort biais ou non avec la fonction de scipy $chi2\_contingency()$ qui renvoie les valeurs χ^2 et $p - value$.

$$\chi^2 = 37.99 \mid p - value = 0.0000$$

Avec de telles valeurs, il semble fortement probable qu'il existe un biais sur l'âge et le genre combinés, qui devra être étudié plus en profondeur et corrigé par la suite.

Nous avons ensuite regardé la répartition des Vrais Positifs et Faux Positifs par âge et genre grâce aux fonctions $show\_tpr()$ et $show\_fpr()$ pour voir plus en détail le potentiel biais.

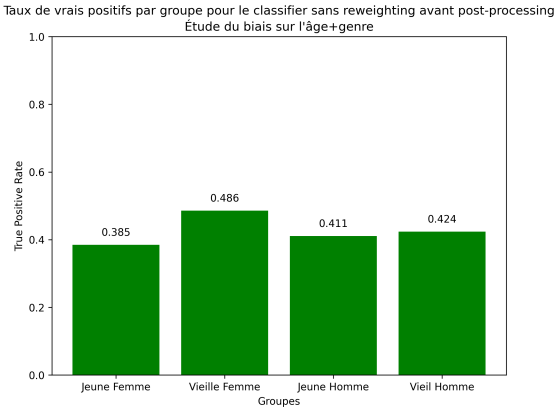

FIGURE 8 – Proportion de Vrais Positifs en fonction de l'âge et du genre

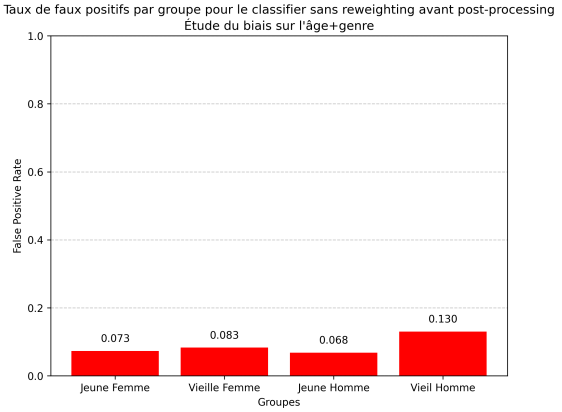

FIGURE 9 – Proportion de Faux Positifs en fonction de l'âge et du genre

$$Différence\ TPR : 0.101 \mid Différence\ FPR : 0.062$$

On peut remarquer que le modèle sans correction des biais a une tendance à prédire plus souvent un cas positif pour une personne âgée, et que les prédictions sur les femmes ont l'air également un peu meilleures que pour un homme, correspondant donc bien à l'assemblage des deux biais étudiés séparément.

3 Pré-processing

3.1 Reweighting par répartition des individus selon leur attribut sensible et leur label

Le premier reweighting implémenté avec la fonction *reweight_by_group_and_label*() affecte à chaque individu un poids selon se répresentation dans l'attribut sensible et son label.

3.1.1 Biais sur l'âge

On donne ci-dessous les poids associés à chaque individu selon son âge et son label.

$$Vieux\ Sain = 1.242 \mid Vieux\ Malade = 1.005 \mid Jeune\ Sain = 0.755 \mid Jeune\ Malade = 1.143$$

On donne maintenant ci-dessous les performances de ce modèle, puis l'analyse en Vrais Positifs et Faux Positifs par catégorie d'âge, en mettant à gauche les résultats de l'ancien modèle (sans reweighting) et à droite les résultats du nouveau modèle.

$$balanced\_accuracy = 0.684 \mid accuracy = 0.693$$

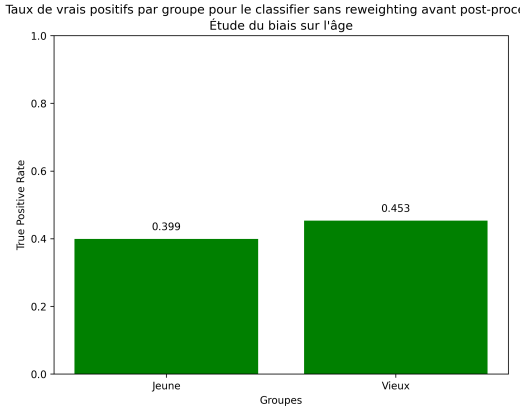

FIGURE 10 – Proportion de Vrais Positifs en fonction de l'âge (avant reweighting)

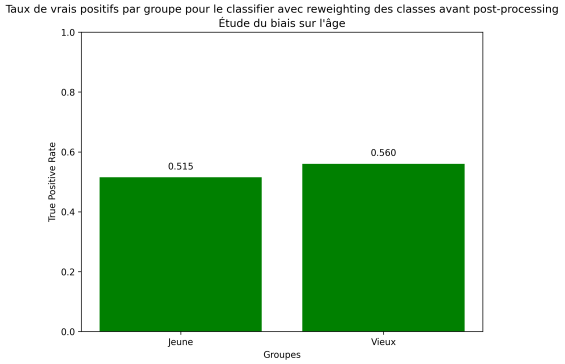

FIGURE 11 – Proportion de Vrais Positifs en fonction de l'âge (après reweighting selon la répartition âge-label)

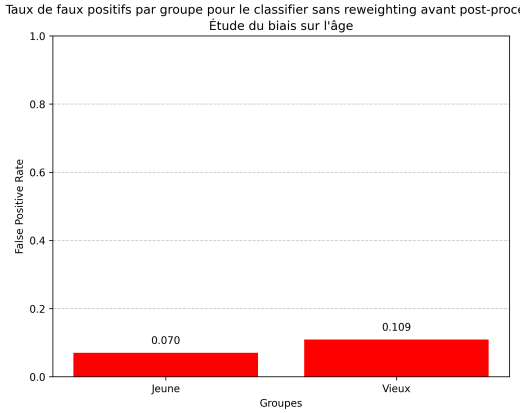

FIGURE 12 – Proportion de Faux Positifs en fonction de l'âge (avant reweighting)

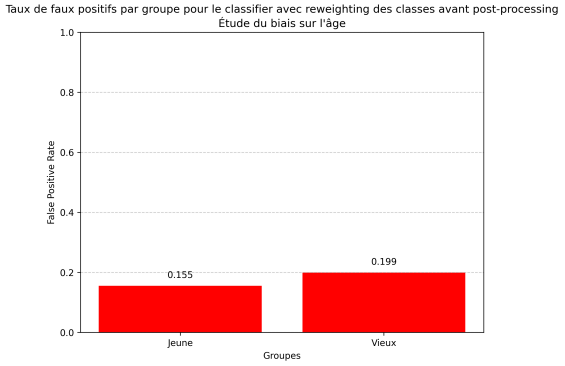

FIGURE 13 – Proportion de Faux Positifs en fonction de l'âge (après reweighting selon la répartition âge-label)

$$Différence\ TPR : 0.045 \mid Différence\ FPR : 0.044$$

On peut remarquer que le reweighting a d'abord beaucoup fait augmenter le nombre de prédictions positives, que ce soit parmis les Vrais Positifs que les Faux Positifs, mais n'a également pas énormément réduit le biais, la différence entre les deux classes en terme de TPR et FPR étant toujours similaire, et le modèle a toujours une forte tendance à plus prédire positif pour une personne âgée.

3.1.2 Biais sur le genre

On donne ci-dessous les poids associés à chaque individu selon son genre et son label.

$$Homme\ Sain = 0.895 \mid Homme\ Malade = 0.992 \mid Femme\ Saine = 0.987 \mid Femme\ Malade = 1.161$$

On donne maintenant ci-dessous les performances de ce modèle, puis l'analyse en Vrais Positifs et Faux Positifs par genre, en mettant à gauche les résultats de l'ancien modèle (sans reweighting) et à droite les résultats du nouveau modèle.

$$balanced\_accuracy = 0.710 \mid accuracy = 0.711$$

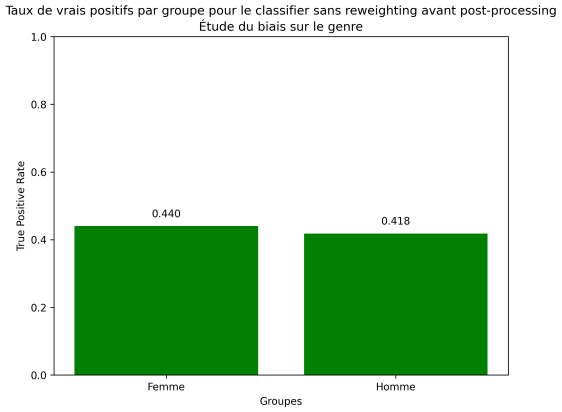

FIGURE 14 – Proportion de Vrais Positifs en fonction du genre (avant reweighting)

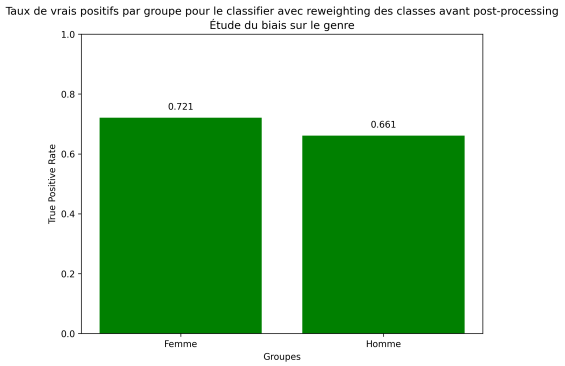

FIGURE 15 – Proportion de Vrais Positifs en fonction du genre (après reweighting selon la répartition genre-label)

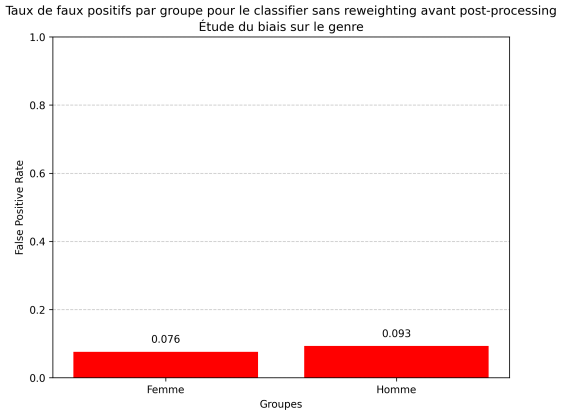

FIGURE 16 – Proportion de Faux Positifs en fonction du genre (avant reweighting)

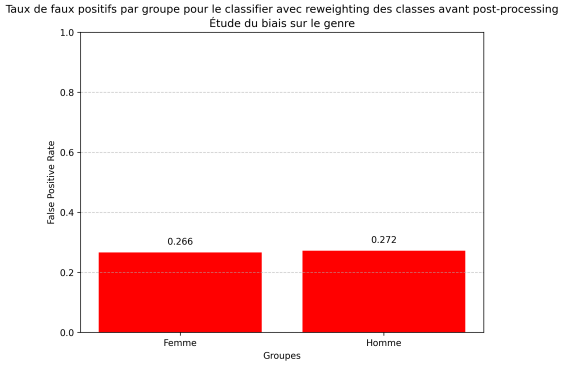

FIGURE 17 – Proportion de Faux Positifs en fonction du genre (après reweighting selon la répartition genre-label)

$$Différence\ TPR : 0.060 \mid Différence\ FPR : 0.006$$

On peut remarquer que le reweighting a d'abord beaucoup fait augmenter le nombre de prédictions positives, que ce soit parmis les Vrais Positifs que les Faux Positifs, mais qu'il a par contre légérement augmenté le biais, la différence entre les deux classes en terme de TPR étant nettement plus haute, mais la différence en terme de FPR étant légèrement plus basse. De plus, le modèle obtient des performances légèrement meilleures mais a maintenant une tendance à plus prédire positif pour une femme.

3.1.3 Biais sur l'âge et le genre

On donne ci-dessous les poids associés à chaque individu selon son âge et genre et son label.

$$Jeune\ Femme\ Malade = 1.267 \mid Vieille\ Femme\ Malade = 1.071 \mid Jeune\ Homme\ Malade = 1.042$$

$$Vieil\ Homme\ Malade = 0.947 \mid Jeune\ Femme\ Saine = 0.759 \mid Vieille\ Femme\ Saine = 1.410$$

$$Jeune\ Homme\ Sain = 0.750 \mid Vieil\ Homme\ Sain = 1.109$$

On donne maintenant ci-dessous les performances de ce modèle, puis l'analyse en Vrais Positifs et Faux Positifs par catégorie d'âge et genre, en mettant à gauche les résultats de l'ancien modèle (sans reweighting) et à droite les résultats du nouveau modèle.

$$balanced\_accuracy = 0.697 \mid accuracy = 0.703$$

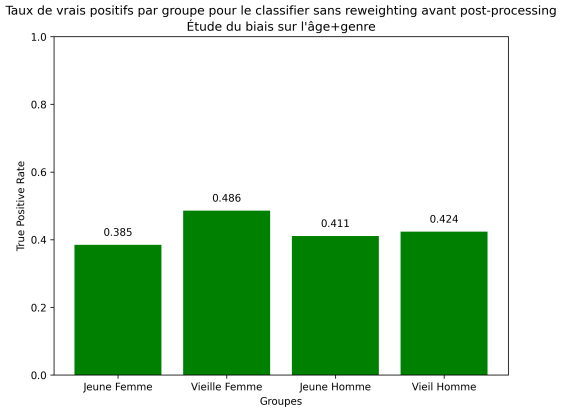

FIGURE 18 – Proportion de Vrais Positifs en fonction de l'âge et du genre (avant reweighting)

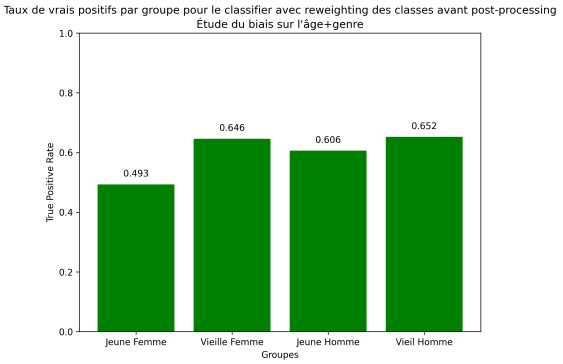

FIGURE 19 – Proportion de Vrais Positifs en fonction de l'âge et du genre (après reweighting selon la répartition âge/genre-label)

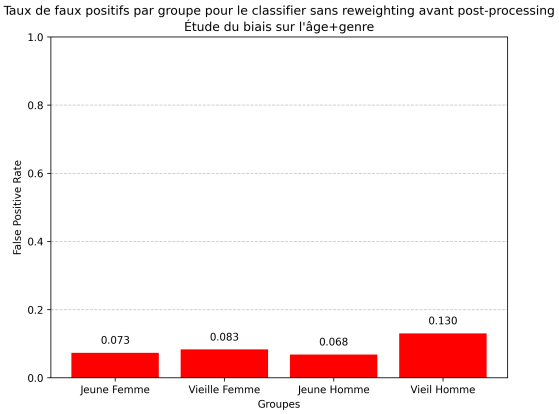

FIGURE 20 – Proportion de Faux Positifs en fonction de l'âge et du genre (avant reweighting)

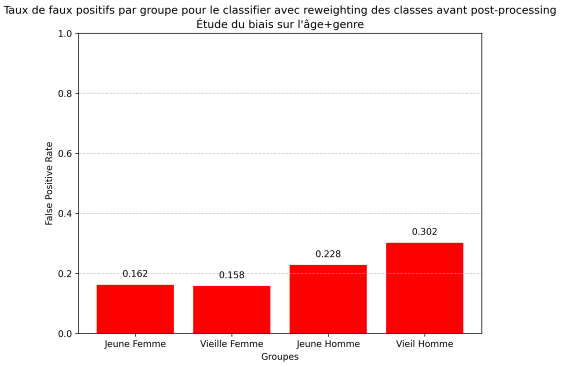

FIGURE 21 – Proportion de Faux Positifs en fonction de l'âge et du genre (après reweighting selon la répartition âge/genre-label)

$$Différence\ TPR : 0.158 \mid Différence\ FPR : 0.144$$

On peut remarquer que le reweighting a d'abord beaucoup fait augmenter le nombre de prédictions positives, que ce soit parmis les Vrais Positifs que les Faux Positifs, mais a augmenté le biais, la différence entre les deux classes que ce soit en terme de TPR et FPR ayant augmenté. Le modèle a toujours une forte tendance à plus prédire positif pour une personne âgée et pour les hommes.

3.2 Reweighting selon la formule de Kamiran-Calders

Le second reweighting implémenté avec la fonction $reweight\_kamiran\_calders()$ affecte à chaque individu un poids selon la formule de Kamiran-Calders.

3.2.1 Biais sur l'âge

On donne ci-dessous les poids associés à chaque individu selon son âge et son label.

$$Vieux\ Sain = 1.190 \mid Vieux\ Malade = 0.846 \mid Jeune\ Sain = 0.884 \mid Jeune\ Malade = 1.175$$

On donne maintenant ci-dessous les performances de ce modèle, puis l'analyse en Vrais Positifs et Faux Positifs par catégorie d'âge, en mettant à gauche les résultats de l'ancien modèle (sans reweighting) et à droite les résultats du nouveau modèle.

$$balanced\_accuracy = 0.714 \mid accuracy = 0.713$$

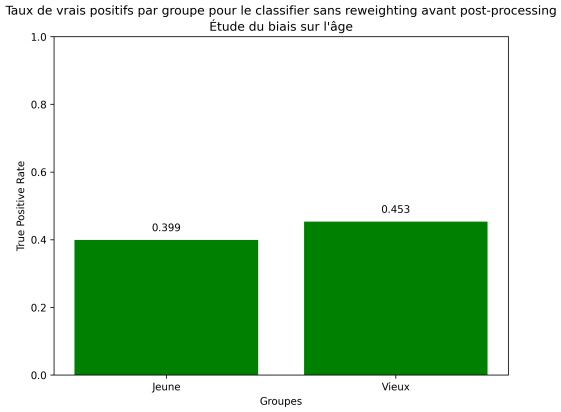

FIGURE 22 – Proportion de Vrais Positifs en fonction de l'âge (avant reweighting)

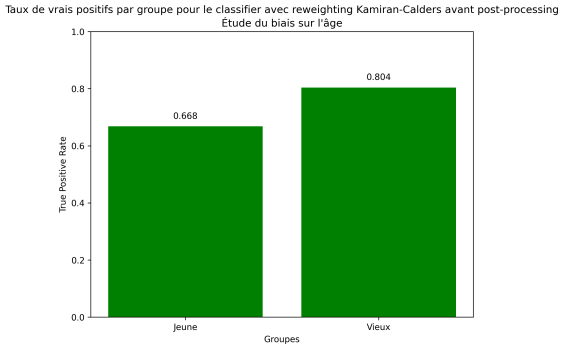

FIGURE 23 – Proportion de Vrais Positifs en fonction de l'âge (après reweighting selon la formule de Kamiran-Calders)

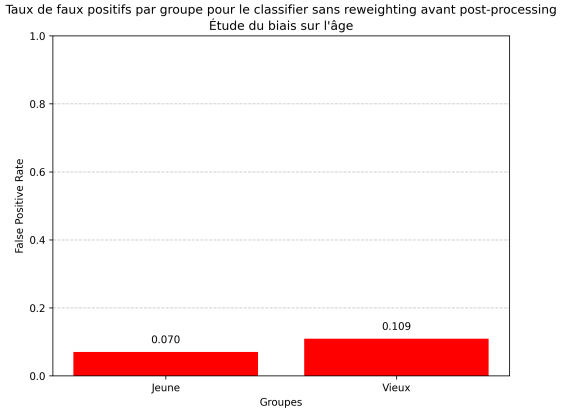

FIGURE 24 – Proportion de Faux Positifs en fonction de l'âge (avant reweighting)

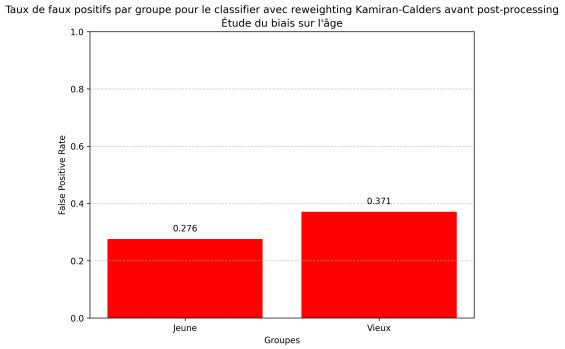

FIGURE 25 – Proportion de Faux Positifs en fonction de l'âge (après reweighting selon la formule de Kamiran-Calders)

$$Différence\ TPR : 0.137 \mid Différence\ FPR : 0.095$$

On peut remarquer que le reweighting a d'abord beaucoup fait augmenter le nombre de prédictions positives, que ce soit parmis les Vrais Positifs que les Faux Positifs, mais a beaucoup augmenté le biais, la différence entre les deux classes en terme de TPR et FPR ayant augmenté. Le modèle a toujours une forte tendance à plus prédire positif pour une personne âgée mais a par contre de meilleures performances globales.

3.2.2 Biais sur le genre

On donne ci-dessous les poids associés à chaque individu selon son genre et son label.

$$Homme\ Sain = 1.013 \mid Homme\ Malade = 0.985 \mid Femme\ Saine = 0.985 \mid Femme\ Malade = 1.017$$

On donne maintenant ci-dessous les performances de ce modèle, puis l'analyse en Vrais Positifs et Faux Positifs par genre, en mettant à gauche les résultats de l'ancien modèle (sans reweighting) et à droite les résultats du nouveau modèle.

$$balanced\_accuracy = 0.709 \mid accuracy = 0.699$$

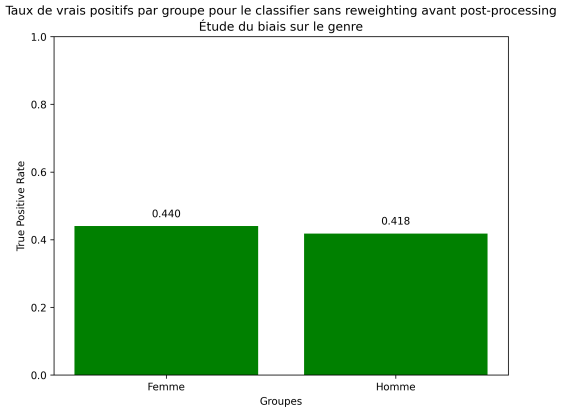

FIGURE 26 – Proportion de Vrais Positifs en fonction du genre (avant reweighting)

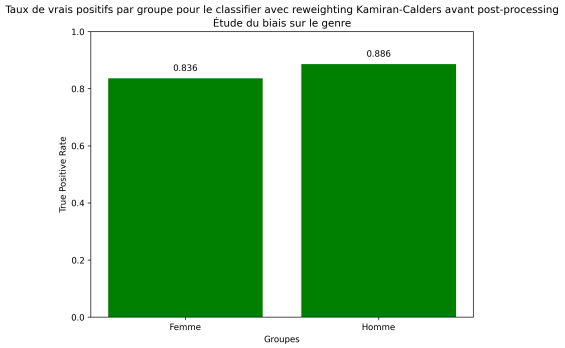

FIGURE 27 – Proportion de Vrais Positifs en fonction du genre (après reweighting selon la formule de Kamiran-Calders)

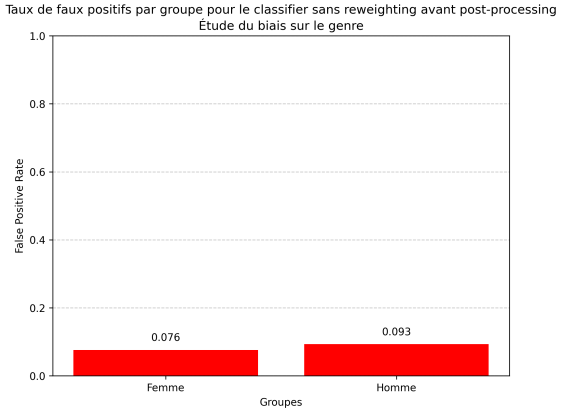

FIGURE 28 – Proportion de Faux Positifs en fonction du genre (avant reweighting)

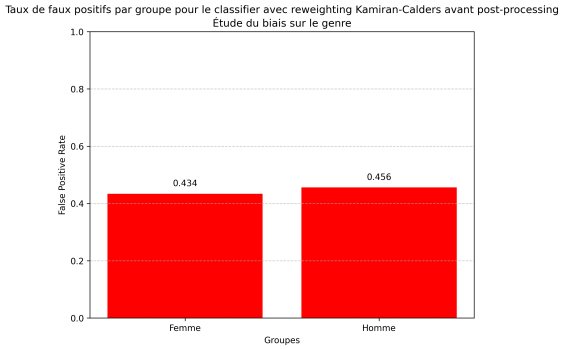

FIGURE 29 – Proportion de Faux Positifs en fonction du genre (après reweighting selon la formule de Kamiran-Calders)

$$Différence\ TPR : 0.050 \mid Différence\ FPR : 0.022$$

On peut remarquer que le reweighting a d'abord beaucoup fait augmenter le nombre de prédictions positives, que ce soit parmis les Vrais Positifs que les Faux Positifs, mais qu'il a par contre légérement augmenté le biais, la différence entre les deux classes en terme de TPR et de FPR ayant augmenté. De plus, le modèle obtient des performances légèrement meilleures mais a maintenant une tendance à plus prédire positif pour un homme.

3.2.3 Biais sur l'âge et le genre

On donne ci-dessous les poids associés à chaque individu selon son âge et genre et son label.

$$Jeune\ Femme\ Malade = 1.247 \mid Vieille\ Femme\ Malade = 0.823 \mid Jeune\ Homme\ Malade = 1.116$$

$$Vieil\ Homme\ Malade = 0.866 \mid Jeune\ Femme\ Saine = 0.852 \mid Vieille\ Femme\ Saine = 1.234$$

$$Jeune\ Homme\ Sain = 0.916 \mid Vieil\ Homme\ Sain = 1.157$$

On donne maintenant ci-dessous les performances de ce modèle, puis l'analyse en Vrais Positifs et Faux Positifs par catégorie d'âge et genre, en mettant à gauche les résultats de l'ancien modèle (sans reweighting) et à droite les résultats du nouveau modèle.

$$balanced\_accuracy = 0.701 \mid accuracy = 0.695$$

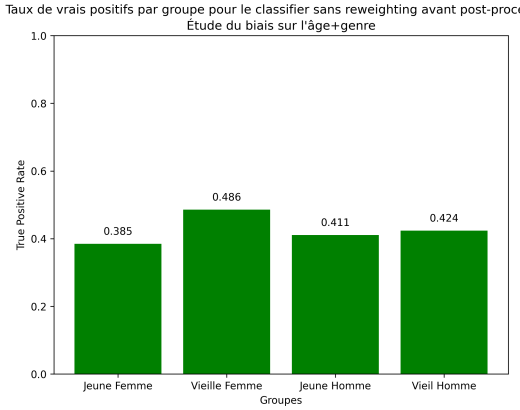

FIGURE 30 – Proportion de Vrais Positifs en fonction de l'âge et du genre (avant reweighting)

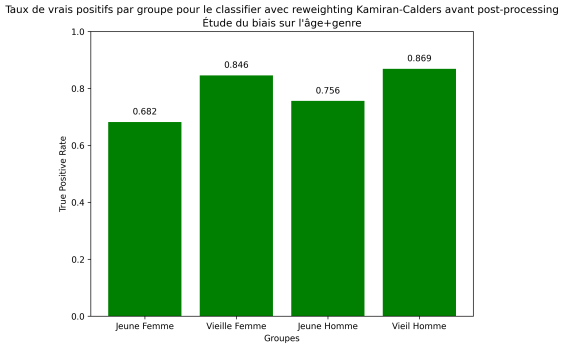

FIGURE 31 – Proportion de Vrais Positifs en fonction de l'âge et du genre (après reweighting selon la formule de Kamiran Calders)

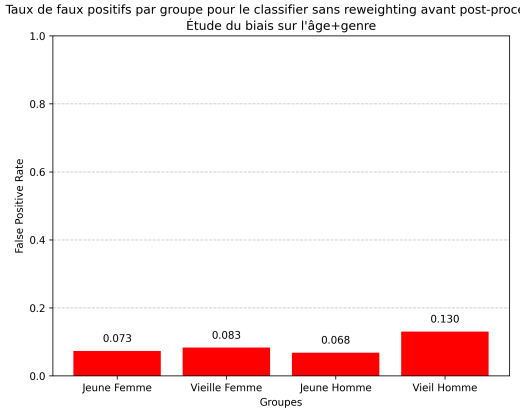

FIGURE 32 – Proportion de Faux Positifs en fonction de l'âge et du genre (avant reweighting)

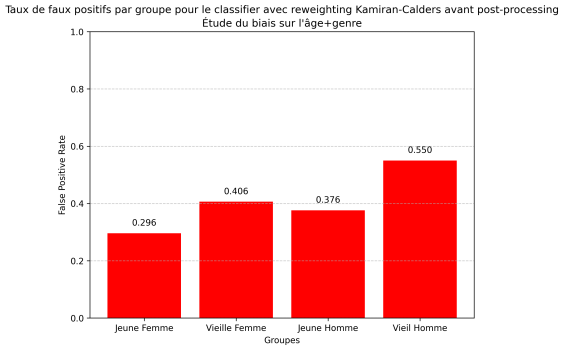

FIGURE 33 – Proportion de Faux Positifs en fonction de l'âge et du genre (après reweighting selon la formule de Kamiran Calders)

$$Différence\ TPR : 0.186 \mid Différence\ FPR : 0.255$$

On peut remarquer que le reweighting a d'abord beaucoup fait augmenter le nombre de prédictions positives, que ce soit parmis les Vrais Positifs que les Faux Positifs, mais a augmenté le biais, la différence entre les deux classes que ce soit en terme de TPR et FPR ayant augmenté. Le modèle a toujours une forte tendance à plus prédire positif pour une personne âgée et pour les hommes.

4 Post-Processing

4.1 Reject Option

Le premier post-processing implémenté avec la fonction *reject_option*() inverse la prédiction d'un individu du groupe défavorisé en TPR et en FPR si le taux de confiance de la prédiction est inférieur à un certain seuil. Le seuil a été entraîné manuellement pour minimiser le biais.

Pour obtenir ce taux de confiance, nous avons légèrement modifié la fonction *preds_todf*() pour qu'elle ajoute au fichier *preds.csv* une colonne $confidence$, où les valeurs sont comprises entre 0.5 et 1, et où une valeur proche de 0.5 signifie une forte incertitude sur la prédiction.

On donne dans ce rapport les résultats de cette méthode en se servant du classifier sans reweighting pour éviter de se répéter, mais les résultats et analyses de cette méthode sur chaque classifier (avec les deux méthodes de pré-processing) sont disponibles dans les fichiers *.txt* et les plots associés sont disponibles dans le dossier *plots*.

4.1.1 Biais sur l'âge

On donne ci-dessous les performances de ce modèle, puis l'analyse en Vrais Positifs et Faux Positifs par catégorie d'âge, en mettant à gauche les résultats de l'ancien modèle (sans post-processing) et à droite les résultats du nouveau modèle.

$$balanced\_accuracy = 0.676 \mid accuracy = 0.691$$

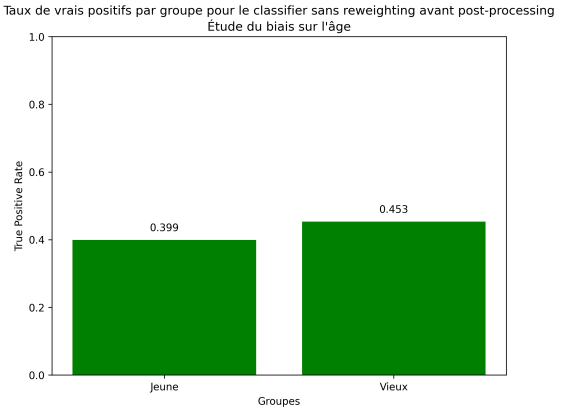

FIGURE 34 – Proportion de Vrais Positifs en fonction de l'âge (avant post-processing)

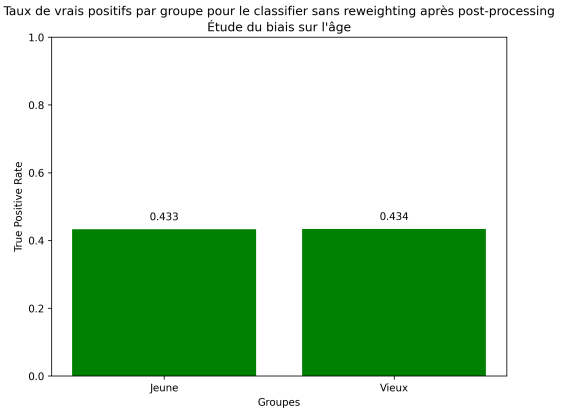

FIGURE 35 – Proportion de Vrais Positifs en fonction de l'âge (après application de Reject Option)

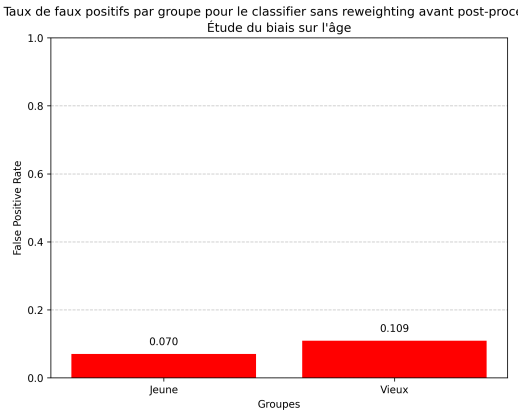

FIGURE 36 – Proportion de Faux Positifs en fonction de l'âge (avant post-processing)

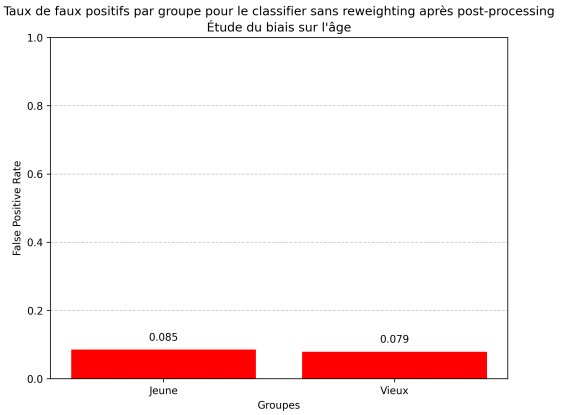

FIGURE 37 – Proportion de Faux Positifs en fonction de l'âge (après application de Reject Option)

$$Différence\ TPR : 0.001 \mid Différence\ FPR : 0.005$$

On peut remarquer que le post-processing a presque totalement enlevé le biais et que le modèle semble maintenant presque parfaitement équilibré.

4.1.2 Biais sur le genre

On donne maintenant ci-dessous les performances de ce modèle, puis l'analyse en Vrais Positifs et Faux Positifs par genre, en mettant à gauche les résultats de l'ancien modèle (sans post-processing) et à droite les résultats du nouveau modèle.

$$balanced\_accuracy = 0.677 \mid accuracy = 0.692$$

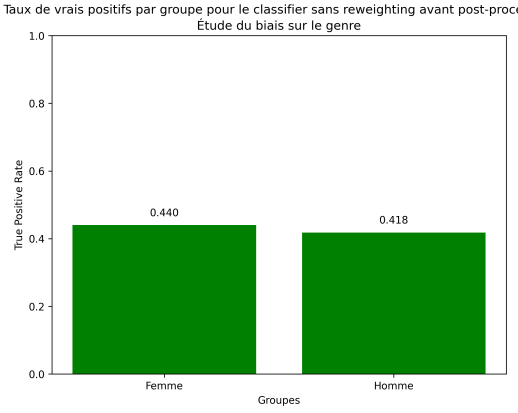

FIGURE 38 – Proportion de Vrais Positifs en fonction du genre (avant post-processing)

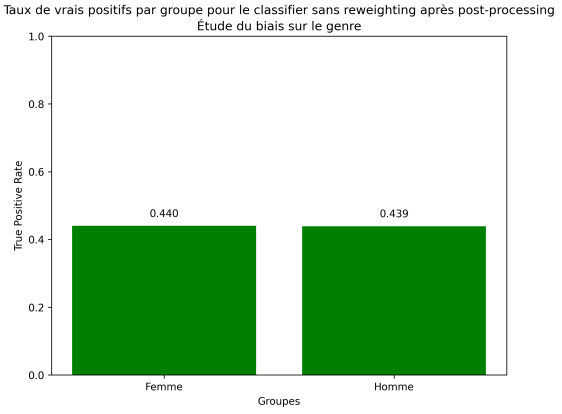

FIGURE 39 – Proportion de Vrais Positifs en fonction du genre (après application de Reject Option)

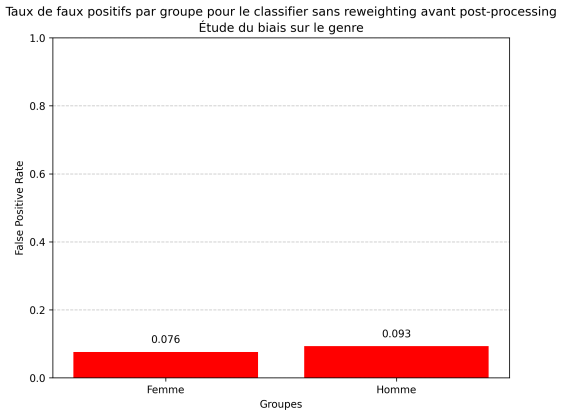

FIGURE 40 – Proportion de Faux Positifs en fonction du genre (avant post-processing)

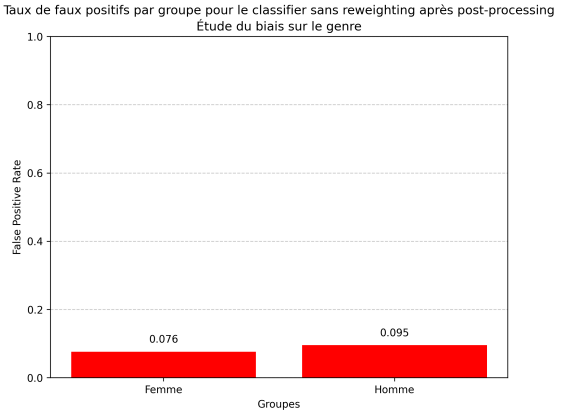

FIGURE 41 – Proportion de Faux Positifs en fonction du genre (après application de Reject Option)

$$Différence\ TPR : 0.000 \mid Différence\ FPR : 0.019$$

On peut remarquer que le post-processing a énormément réduit le biais et que le modèle semble maintenant presque parfaitement équilibré, avec un très léger biais envers les hommes.

4.1.3 Biais sur l'âge et le genre

On donne maintenant ci-dessous les performances de ce modèle, puis l'analyse en Vrais Positifs et Faux Positifs par catégorie d'âge et genre, en mettant à gauche les résultats de l'ancien modèle (sans post-processing) et à droite les résultats du nouveau modèle.

$$balanced\_accuracy = 0.674 \mid accuracy = 0.690$$

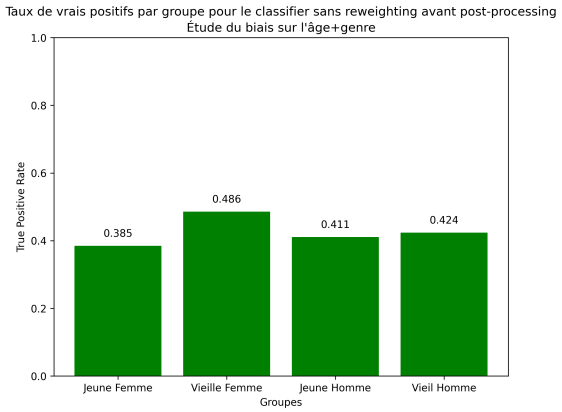

FIGURE 42 – Proportion de Vrais Positifs en fonction de l'âge et du genre (avant post-processing)

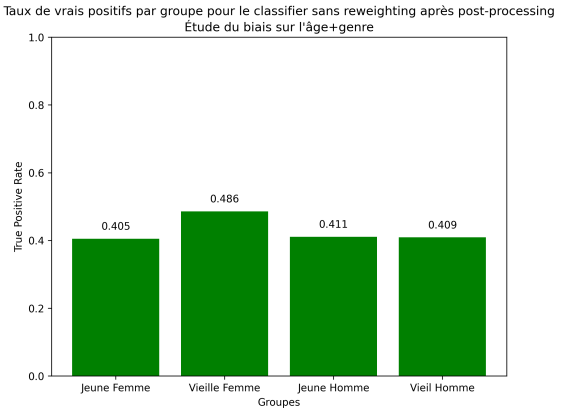

FIGURE 43 – Proportion de Vrais Positifs en fonction de l'âge et du genre (après application de Reject Option)

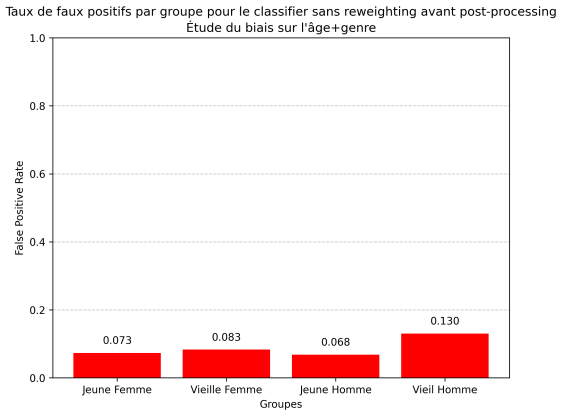

FIGURE 44 – Proportion de Faux Positifs en fonction de l'âge et du genre (avant post-processing)

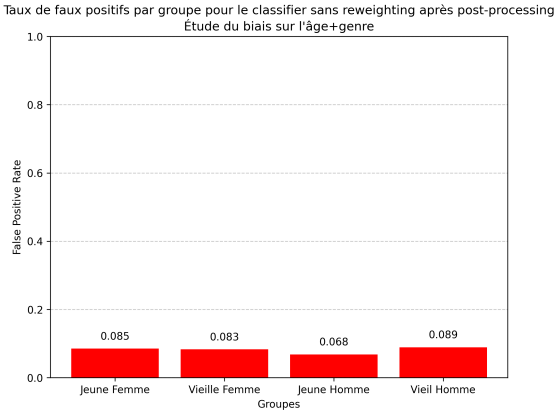

FIGURE 45 – Proportion de Faux Positifs en fonction de l'âge et du genre (après application de Reject Option)

$$Différence\ TPR : 0.080 \mid Différence\ FPR : 0.021$$

On peut remarquer que le post-processing a légèrement réduit le biais mais que le modèle semble quand même légèrement discriminer les hommes et les personnes jeunes.

4.2 Equalized Odds

Le second post-processing implémenté avec la fonction *equalized_odds_postprocessing*() inverse avec une certaine probabilité la prédiction d'un individu du groupe défavorisé d'après les métriques *Equalized Odds* si le taux de confiance de la prédiction est inférieur à un certain seuil.

Cette méthode nécessitant deux hyper-paramètres à entrainer (le seuil de confiance à considérer et la probabilité d'inverser une prédiction), et étant dépendante du hasard, son entraînement et son analyse est plus complexe et n'est pas présentée dans le rapport. On a cependant noté une baisse du biais sur quelques essais, mais qualitativement moins bon qu'avec *Reject Option*.

5 Conclusion

Pour conclure, le biais le plus important portait sur l'âge, le genre étant très peu discriminé. Avec différentes méthodes appliquées, on a pu réduire le biais, mais également augmenter la tendance à prédire positif ou négatif, résultant dans plus de Vrais Positifs, mais en même temps plus de Faux Positifs.

Les méthodes de post-processing se sont montrées bien plus efficaces pour réduire le biais, tandis que les méthodes de pré-processing se sont quant à elles montrées plus efficaces pour augmenter les prédictions positives (ce qui peut être bien dans notre cas tant que cela reste raisonnable, puisque un positif subira des examens supplémentaires tandis qu'un négatif sera renvoyé chez lui).

Le meilleur modèle en terme de Fairness semble être celui n'utilisant pas de reweighting et appliquant l'algorithme de *Reject Option*.

Les modèles utilisant un reweighting et appliquant l'algorithme de *Reject Option* semblent cependant meilleurs pour maximiser le taux de Vrais Positifs sans pour autant trop augmenter le taux de Faux Positifs. Le reweighting avec la formule de $Kamiran - Calders$ est celui qui augmente le plus le taux de prédictions positives, tandis que celui utilisant un reweighting en fonction de la répartition des individus selon leur attribut sensible et leur label semble plus modéré dans l'augmentation des prédictions positives.